

## Lecture 3 – Alignment

### I. Introduction.

For sequence data, the process of generating an alignment establishes positional homologies, that is, the identification of homologous phylogenetic characters.

Because there are only five possible character states for each character and the states are common across all the characters, establishing character homologies can be very challenging.

Furthermore, there is no way to apply some of the criteria that are useful in positing homologies for morphological characters (transitional forms, developmental similarity). Thus, we're forced to rely on positional similarity as determined in our alignments to provide homologous characters.

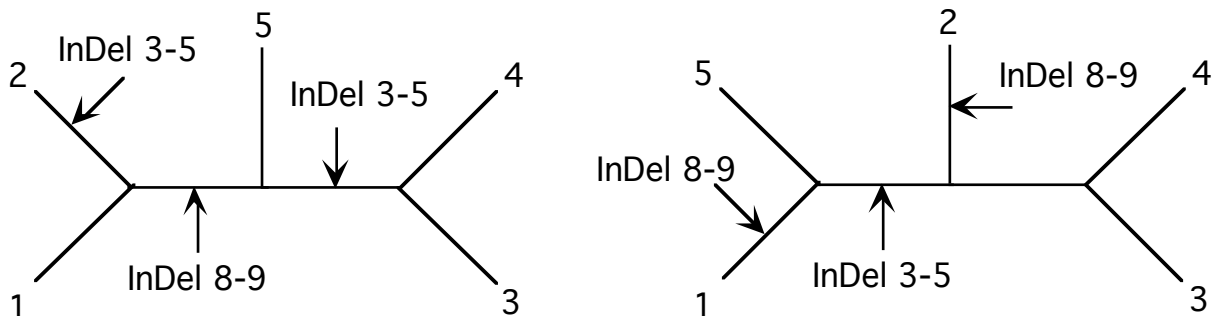
**II. The ideal approach** was discovered 30 years ago. It involves **simultaneous alignment and tree estimation** (Sankoff et al., 1973; *Nature New Biology*, 245:232). This is actually compatible with the ideas of testing morphological hypotheses of homology with phylogenetic analyses we discussed last week.

Alignment is essentially an inference of insertion/deletion events (indels).

Let's think for a minute why alignment depends on the phylogeny (we'll use the example from Chapter 29 in the text).

```
1   ACCGAAT--ATTAGGCTC
2   AC---AT--AGTGGGATC
3   AA---AGGCATTAGGATC
4   GA---AGGCATTAGCATC
5   CACGAAGGCATTGGGCTC
```

So, it looks like there have be two insertion/deletion events, one at positions 3-5 and one at positions 8-9, but there's no tree on which both can evolve only one time.



So really, figuring out where insertions and deletions have occurred requires a historical framework (i.e., a phylogeny), which, as far as we're concerned, is the motivation for collecting sequence data in the first place.

We can think about alignment in terms of an overall score, including a score for substitutions and a cost for gaps: A simple score might look like this:

$$D = s + wg.$$

$s$  is the score for (mis)matches,  $g$  is a gap cost and  $w$  is a weighting factor for gaps.

We'll worry about the details of these values in few minutes, but the ideal approach would be to find that topology that has the best  $D$ , across all possible topologies and all possible alignments (i.e., find the globally optimal combination of topology + alignment).

However, given that phylogenetic estimation from a single alignment is a problem that is so computationally difficult as to require heuristic methods (short cuts), simultaneous alignment is really not feasible.

III. The most widely use heuristic is called **progressive alignment**, and the most commonly used program is **Clustal W**.

A. Overview. Alignment occurs in three steps.

1. A pair-wise alignment is generated for all  $(n^2-n)/2$  pairs of sequences. A pair-wise distance is estimated between each of the pairs and this is entered into a distance matrix.
2. That distance matrix is subjected to an algorithmic tree-building procedure (NJ).
3. This guide tree is used to align progressively more distant pairs of sequences, until all sequences are in the alignment.

B. Pair-wise alignments are usually conducted using Needleman-Wunsch (1970) algorithm.

Take two sequences:

G A A T T C A G T T A (#1)

G G A T C G A (#2)

So  $L_1 = 11$  and  $L_2 = 7$

To illustrate, we can assume a simple scoring scheme (i.e., set of alignment parameters).



We can therefore fill in the matrix:

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5	5
A	0	1	2	3	3	3	3	4	5	5	5	6

Now, we trace back from the bottom right to the top left, following the path with the highest score.

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	<b>1</b>	1	1	1	1	1	1	1	1	1	1
G	0	<b>1</b>	1	1	1	1	1	1	2	2	2	2
A	0	1	<b>2</b>	<b>2</b>	2	2	2	2	2	2	2	3
T	0	1	2	2	<b>3</b>	<b>3</b>	3	3	3	3	3	3
C	0	1	2	2	3	3	<b>4</b>	<b>4</b>	4	4	4	4
G	0	1	2	2	3	3	4	4	<b>5</b>	<b>5</b>	<b>5</b>	5
A	0	1	2	3	3	3	3	4	5	5	5	<b>6</b>

This gives the alignment:

```

G - A A T T C A G T T A
|   |   |   |   |
G G A - T - C - G - - A

```

Which has a score of 6 (there are six matches). **All paths with a score of 6** represent optimal pair-wise alignments, with the alignment parameters that we assumed.

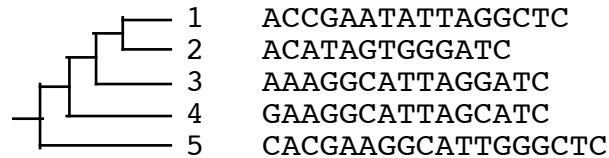
So in a multiple alignment, this is done for all  $(n^2-n)/2$  pair-wise comparisons of sequences. These pairwise alignments are used to estimate pair-wise distances, which are entered into a pair-wise matrix.

C. This matrix is then subjected to an algorithmic method of generating a tree.

The Neighbor Joining method is usually used in Clustal W. This is a star decomposition approach, so it's fast.

- D. This NJ tree is used to build a multiple alignment progressively, first aligning the closest sequences on the guide tree, and progressively aligning sequences that are more distant.

#### Guide Tree



Following the Guide Tree, 1 & 2 are aligned.

```

1  ACCGAATATTAGGCTC
2  AC---ATAGTGGGATC
  
```

Again, following the guide tree, we align sequence 3 to the fixed 1/2 alignment.

```

1  ACCGAAT--ATTAGGCTC
2  AC---AT--AGTGGGATC
3  AA---AGGCATTAGGATC
  
```

We continue to follow the guide tree building a multiple alignment by progressive pairwise alignments.

```

1  ACCGAAT--ATTAGGCTC
2  AC---AT--AGTGGGATC
3  AA---AGGCATTAGGATC
4  GA---AGGCATTAGCATC
5  CACGAAGGCATTGGGCTC
  
```

#### E. Elaborations and refinements:

- A substitution matrix is used to calculate the scores for substitutions. This allows for biochemically similar amino acid substitutions to cost less than radical ones (for protein alignments) or transitional substitutions to cost less than transversional substitutions (for DNA alignments).
- Gap costs can be elaborated so that there is a separate cost associated with opening a new gap (GOP) vs, extending a gap (GEP).

- Each sequence is weighted according to how different it is from the other sequences. This accounts for the case where one specific subfamily is over represented in the data set.
- Position-specific gap-open penalties are modified according to residue type using empirical observations in a set of alignments based on 3D structures. In general, hydrophobic residues have higher gap penalties than hydrophilic, since they are more likely to be in the hydrophobic core, where gaps should not occur.
- A big computational cost is doing all the pair-wise Needleman-Wunsch alignments to build a guide tree. MUSCLE (Edgar, 2004) generates a guide tree from an alternative approach. It uses distance based on frequency of shared k-mer words from all pairs of sequences to generate a guide tree.

The Euclidian Distance between a pair of sequences,  $i$  &  $j$  (for  $k = 5$ ) :

$$d_{ij} = \left\{ \sum_{x=1}^{1024} (f_{xi} - f_{xj})^2 \right\}^{1/2}$$

Then it does a progressive alignment, as above. It then estimates a quick and dirty tree from the initial alignment and iterates another round of progressive alignments. This continues until the same tree recurs in successive rounds.

- MAFFT (Kato, 2002) takes a similar approach, but permits the incorporation of biochemical properties of amino acids in doing protein sequence alignments.
- Golubchick et al. (2007) evaluate these and a few other approaches.
- Gblocks (Castresana, 2000) automates elimination of ambiguously aligned regions, which has been shown to dramatically improve phylogeny estimation (Talavera and Castresana, 2007).

## F. Models in Alignment

One problem is that the final alignment is dependent on the parameters, and the optimum parameter values depend on unknown processes of evolution (i.e., insertion rate, deletion rate, and rates of various substitution types).

The kicker is that we need a good alignment to estimate these. This is because, as discussed in the beginning of this lecture, alignment and historical information are not independent of each other.

A small but growing body of work attempts to model this so that we can simultaneously optimize parameters and alignments.

These are extremely computationally intensive. Although simultaneous optimization is superior in theory, it's unclear if the approximations that are currently in use and in-development, will actually produce substantially better alignments than progressive alignment.

Furthermore, all computational methods have a difficult time producing alignments that maintain secondary structure appropriately, especially for rRNA sequences.

A very abbreviated list of relevant references:

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17:540-552.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucl. Acids Res.* 32:1792-97.

Feng, D.-F. and R. F. Doolittle. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* 25: 351-360.

Fleißner, R., D. Metzler & A. von Haeseler. 2005. Simultaneous Statistical Multiple Alignment and Phylogeny Reconstruction. *Syst. Biol.* 54:548-561.

Golubchik, T., M. J. Wise, S. Easteal, & L. S. Jermin. 2007. Mind the Gaps: Evidence of Bias in Estimates of Multiple Sequence Alignments. *Mol. Biol. Evol.* 24:2433-2442.

Hein, J. J., L. Jensen, & C. N. S. Pedersen. 2003. Recursions for statistical multiple alignment. *PNAS* 100:14950-14965.

Katoh, K., K. Misawa, K. Kuma, and T. Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucl. Acids Res.* 30:3059-3036.

Metzler, D., R. Fleißner, A. Wakolbinger, & A. von Haeseler. 2001. Assessing variability by joint sampling of alignments and mutation rates. *J. Mol. Evol.* 53:660-669.

Needleman, S. B. & C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acids sequence of two proteins. *J. Mol. Biol.* 48:443-453.

Smith, T. F. & M. S. Waterman. 1982. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195-197.

Suchard, M. A., & B. D. Redelings. 2006. BALi-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, 22:2047-2048.

Talavera, G., & J. Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56, 564-577.

Thompson, J.D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673-80.

Wheeler, W. 2001. Homology and the optimization of DNA sequence data. *Cladistics*, 17:S3-S11.